Abstract

Imaging Mass Spectrometry (IMS) is a powerful molecular imaging technology that enables the direct analysis of the spatial distribution of biomolecules in a tissue section. This technology allows for the monitoring of thousands of molecules throughout the tissue in a single experiment, and enables comparison of the biomolecular content between various areas in the tissue, ranging from lowmass metabolites and lipids to high-mass biomacromolecules such as proteins.

IMS requires no hypothesis on possible target molecules beforehand as no labeling of target molecules is involved at any point, making IMS an immensely valuable technology for explorative research. An IMS analysis from a single examined tissue slice, however, also leads to file sizes of several gigabytes and more recently even terabytes of complex and high dimensional data. Manual exploration of these data is becoming increasingly infeasible and in order to harness the full potential of these data, there is a strong need for computational methods that can extract the valuable information that is captured in these mountains of data. It is the development of these computational methods that is the focus of this work.

In this thesis we have concentrated on three computational challenges in particular: feature selection and dimensionality reduction for IMS data, differential analysis of IMS data from multiple experiments, and data fusion of IMS data with additional data sources.

In the feature selection and dimensionality reduction for IMS data, we use the Discrete Wavelet Transform (DWT) as a dimensionality reduction tool that can achieve considerable reduction of the original data size, while keeping all the information in the IMS spectra intact. We accomplished this by discarding only those variables that contain noise, and which take up unnecessary storage space and hamper computational analysis. We were able to further improve on previous work in this field by incorporating spatial information, a key feature of IMS data, into the selection procedure that differentiates between noise and true biological signal. We demonstrate two different ways of incorporating this spatial information, namely global and local spatial information, and show that both methods can improve feature selection over traditional methods. The new methods show a better retention of features that contain true signal and can provide improved data compression and dimensionality reduction by discarding more features that are due to noise.

In the second part of this work we focus on the challenge of computationally comparing IMS data from multiple tissue slices. Performing such a comparison is far from trivial due to the large heterogeneity and the large size of IMS datasets.

We therefore introduce a method called Group Independent Component Analysis (GICA) to perform this comparison. This method allows for an in-depth comparison of a large number of IMS datasets without the need for peak picking, extracting features from the full profile IMS data at the level of an individual tissue sample. By performing the pattern extraction at the level of a single tissue, GICA emphasizes small, tissue-specific features which could be lost in a simultaneous analysis over all tissues. Correct operation of the algorithm is demonstrated using an artificially generated dataset containing 3 different disease classes. This case study shows that the algorithm allows for near-perfect retrieval of the differential components between tissues in the correct locations. Subsequently, this technique is applied to two real-life IMS datasets, a case study of amyotrophic lateral sclerosis (ALS) in mouse spine and one of retinal degeneration disease in mouse retina. Application of the GICA algorithm to these datasets results in retrieval of multiple differential biomolecular patterns between the different tissue classes, which are expressed in biologically relevant areas for the diseases under study.

The third and final topic is the data fusion of IMS data with additional data sources. Combining IMS data with external data sources can create richer datasets and enable deeper insights into the IMS data. We could demonstrate this by linking IMS data directly to an anatomical atlas, namely the Allen Mouse Brain Atlas (AMBA). This integrates anatomical insights, an important resource that many human interpretations will naturally rely upon, with the IMS data. We spatially map IMS data to the atlas using non-rigid image registration techniques and use the established link to investigate spatial correlations between ion images and the anatomical structures of the atlas, gaining basic insights into the

relationships between the two. Subsequently we move beyond simple correlation and use the atlas information to provide automated anatomical interpretation of the ion images in the IMS data. This challenge is solved as a convex optimization problem that deconstructs ion distributions as combinations of known anatomical structures. We demonstrate that establishing a link between an IMS experiment and an anatomical atlas can serve as an important accelerator both for human and machine-guided exploration of IMS experiments.